Monitoring the Gender Gap with Wikidata Human Gender Indicators

Maximilian Klein
GroupLens Research
Dept. of Computer Science
University of Minnesota
max@notconfusing.com

Harsh Gupta
Indian Institute of Technology
Kharagpur
mail@hargup.in

Vivek Rai
Indian Institute of Technology
Kharagpur
vivekraiiitkgp@gmail.com

Piotr Konieczny
Hanyang University
piokon@post.pl

Haiyi Zhu
GroupLens Research
Dept. of Computer Science
University of Minnesota
haiyi@cs.umn.edu

ABSTRACT

The gender gap in Wikipedia's content, specifically in the representation of women in biographies, is well-known but has been difficult to measure. Furthermore the impacts of efforts to address this gender gap have received little attention. To investigate we utilise Wikidata, the database that feeds Wikipedia, and introduce the "Wikidata Human Gender Indicators" (WHGI), a free and open source, longitudinal, biographical dataset monitoring gender disparities across time, space, culture, occupation and language. Through these lenses we show how the representation of women is changing along 11 dimensions. Validations of WHGI are presented against three exogenous datasets: the world's historical population, "traditional" gender-disparity indices (GDI, GEI, GGGI and SIGI), and occupational gender according to the US Bureau of Labor Statistics. Furthermore, to demonstrate its general use in research, we revisit previously published findings on Wikipedia's gender bias that can be strengthened by WHGI.

CCS Concepts

•**Human-centered computing** → **Empirical studies in collaborative and social computing;** Computer supported cooperative work; Wikis;

Keywords

Gender Disparities, Wikipedia, Wikidata, Biographical Database

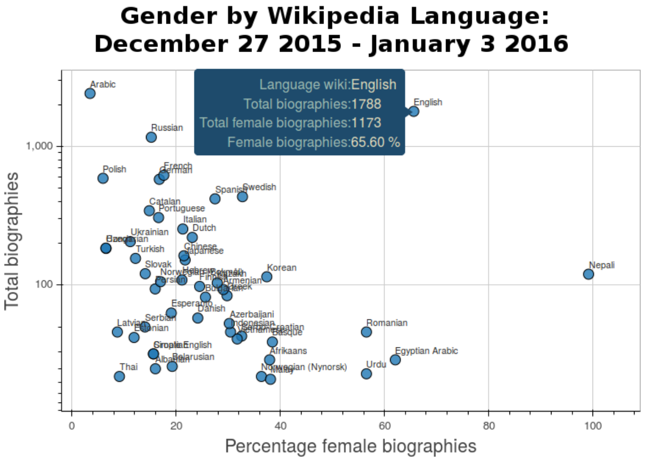


Figure 1: Example of monitoring the changes in biographies of women for Wikipedia languages in the period December 27th 2015 - January 3rd 2016. We can highlight that English Wikipedia increased 1,788 biographies, 65% about women, while Nepali Wikipedia increased by 120 biographies, 119 were about women.

1. INTRODUCTION

Gender inequality is a long-standing social problem which affects many aspects of society. Worldwide, cultural ideologies have created scenarios which make women more prone to health issues [28]. Likewise in education, attitudes create a systemic gender bias in opportunity [13]. And, famously, incomes for identical jobs are lower for women [2].

Statistical gender indicators are critically important to understanding gender inequality [17]. Many indicators (single measures) and indices (compound measures) have been proposed, such as The Gender Development Index from the United Nations Development and the Global Gender Gap Index from the World Economic Forum. Still there are some features lacking among the current gender indicators. First, the most frequent ones are released annually, which does not allow for fine-grained analysis. Also they measure only recent history, which limits analysis of the past. And lastly they are not open source, which means verification and remixing is more difficult.

At the same time Wikipedia's *editor* gender gap has come to light as deeply gender-disparate, and has received media attention for it [3]. Responses, however, are not focused solely on editors, but also *biographical* coverage. Gender-focused Wikipedia editing communities create and improve biographies to combat systemic bias,

OpenSym '16 Berlin, Germany

ISBN 978-1-4503-4451-7/16/08...$15.00

DOI: http://dx.doi.org/10.1145/2957792.2957798

but their large-scale effect is unknown. Luckily, new capabilities to record gender via Wikidata, the database that feeds Wikipedia, make encyclopedia-scale descriptions of gender increasingly tractable and accurate.

We introduce a weekly-updated, all-recorded-history level, open source dataset of the gender of humans in Wikidata. "Wikidata Human Gender Indicators" (WHGI) is a gender disparity indicator, with 11 dimensions of time, space, culture, occupation and language. WHGI is open data and available for download at http://wigi.wmflabs.org/, where we also display visualizations.

Considering the work in indicator research, the Wikipedia Gender Gap, and the introduction of Wikidata, we ask the following research questions:

- **RQ1** How do gender disparity indicators based on the humans in Wikidata reflect real world gender disparities?
- **RQ2** How well does WHGI correlate with existing efforts to close Wikipedia's gender gap?

The rest of this paper is organized around the our research questions. We begin by describing the format of Wikidata and statistics of humans contained in it. Next, we propose WHGI as a gender disparity indicator with 11 dimensions. Then to validate we present three measures utilizing ground truths from the US Census Bureau, Bureau for Labor Statistics, and United Nations Development Program. We also illustrate how Wikipedia editing communities can use our dataset to chart their progress, and help their cause under the philosophy of "what gets measured, gets fixed." Finally, we show how WHGI can be used by researchers to better understand gender disparity phenomenon and other content improvement movements, in ways traditional indicators cannot.

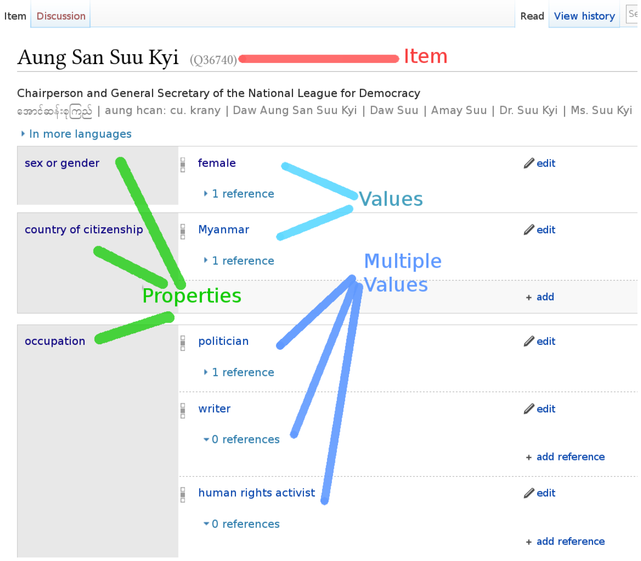
2. RELATED WORK

The thread of research on Wikipedia's gender biases has grown since the finding that Wikipedia's editors are largely not women [8] [19], ranging from only 13% to 16% [14]. This imbalance has been attributed to at least an internet skills gap [12] and the editing community's internal culture [19].

More and more, in addition to investigations of the Wikipedia *editor gender gap* researchers have also been interrogating the character of its *biography gender gap*. Early studies found that Wikipedia excluded notable women more than its counterparts [22]. More recently it was shown that while coverage of women in large Wikipedias is not less than other reference works, the wording with which women are portrayed is different and focuses more on romance and family [25]. Women also tend to be less central in the link graph of Wikipedia [6]. These linguistic and network findings were confirmed by [10], who also showed evidence of stereotyping in metadata. Intersectionally, in terms of socio-economic biases, the level of contributions to a Wikipedia language are associated with the wealth of the country they originate from [21].

Yet in popular mind-share there persists a sentiment that denies that any of this is a problem [5]. Luckily, experiments are showing that awareness of Wikipedia's gender issues is a strategy that can alleviate the problem [15], for which more methods are always needed.

Wikidata, offers new opportunities to analyze culture programmatically. Launched in 2012, Wikidata is designed to host structured data that is *multilingual* (so there is only one edition) and *plural* (can support many competing facts) [24]. These features make Wikidata well-suited for all Wikipedias to collaboratively store facts about the world. If an Italian Wikipedian stores information about the population of ancient Rome, that information is then available to every other Wikipedia with a short code snippet. Every language collaborating together has meant that Wikidata has become a massive free open knowledge-base in its own right, containing over 40 million facts [18].



Item | Discussion | Read | View history

Aung San Suu Kyi (Q36740)

Chairperson and General Secretary of the National League for Democracy
အောင်ဆန်းစုကြည် | aung hcan: cu. krany | Daw Aung San Suu Kyi | Daw Suu | Amay Suu | Dr. Suu Kyi | Ms. Suu Kyi

In more languages

sex or gender: female — 1 reference — edit

country of citizenship: Myanmar — 1 reference — edit — + add

occupation: politician — 1 reference — edit; writer — 0 references — edit — + add reference; human rights activist — 0 references — edit — + add reference

Figure 2: Example Wikidata Human Item of Aung San Suu Kyi

As a knowledge-base, Wikidata is slowly proving its worth for research. Early on Wikidata showed the importance of multilingual Wikipedians in reducing self-focus bias of language editions [11]. Wikidata has also been used to find popular connections between nationalities and occupations [9]. Or take the fact that all human and mouse genes have been imported into Wikidata [20], for an internet-wide community effort to find links between genes, drugs and diseases [1]. All of these tasks would be difficult to do without Wikidata.

3. METHOD

Wikidata is a general database consisting of *items* which are described by *properties* that take on *values*. Our interest is in biographies of people, that is, any item which has the property *instance of* with value *human* [1]. For each human item we find the corresponding values of *gender*, *date of birth*, *date of death*, *place of birth*, *citizenship*, *ethnic group*, *field of work*, and *occupation* [2]. In Figure 2, we illustrate the semantics of a Wikidata Human on the item for Aung San Suu Kyi.

For each of the above eight properties we create an "indicator" by aggregating the dataset on that property, but disaggregating on gender. Take for example the date of birth indicator, it has one row per year found as a date of birth, and one column per gender represented in Wikidata. See a sample excerpt of the date of birth indicator in Table 1, and it's visualization in Figure 3. Notice in particular that not every human with a date of birth has a gender (recorded as *no gender* in our data), and that Wikidata's community has a non-binary view of gender and includes humans which are neither *male* nor *female*. In addition to the eight indi-

[1] As Wikidata is intentionally multilingual, items, properties and values are actually referenced by number. So "instance of: human" is "P31:Q5" in Wikidata terms.

[2] These correspond to Wikidata properties P21, P569, P570, P19, P27, P172, P101, and P106 respectively.

Table 1: An excerpt of the January 3rd 2016 date of birth indicator. Here we can see that the earliest humans in Wikidata are two men born in 4203 BCE, births peak in the 1980's (see Figure 3), and there are thirteen notable 1-year-olds in Wikidata. Also notice the inclusion of non-binary genders in Wikidata, as well as biographies without any gender recorded.

date of birth	no gender	transgender female	genderqueer	kathoey	female	male
4203 (BCE)						2
. . .						
1981	849	1		1	5,042	14,461
1982	861	2			5,132	14,372
1983	864	3			5,078	14,520
1984	830	3	1		5,372	14,558
1985	777	4			5,400	14,664
. . .						
2015	6				4	3

cators made directly from properties, we include three more which feature augmented data. The "language" indicator is based on the Wikipedia languages in which a human is represented (shown in Figure 1), and we include geographic aggregations of citizenship, place of birth and ethnic group into indicators called *culture*[3], and *worldmap* (shown in Figure 4).

3.1 Snapshots

All of our data is derived from the official Wikidata database downloads, which represent a cross-sectional "snapshot" of Wikidata as it was at a specific date. Wikidata releases new snapshots weekly. We re-compute each of the 11 indicators for every new snapshot, and additionally compute the differences that occurred between the newest snapshot and the second-newest. This allows us to monitor activity on Wikidata at a weekly level of granularity. For instance Figure 3 and Figure 4 show the state of the date of birth and country indicators, *for all time*, as of the January 3rd 2016 snapshot, but Figure 1 shows the *changes* of the week between December 27th 2015 - January 3rd 2016.

Therefore, we are also generating a dataset of *weekly changes* which allows us to *monitor* the status of biographies in Wikidata. We can inspect the changes in composition of genders, or date of birth, which can speak to efforts from Wikipedian communities attempting to counter bias in the database.

3.2 Technical Details

Note that for fidelity there is virtually no data-cleaning done, as the point of our project is to display information as faithfully as possible. Our dataset is meant to be used to uncover potential biases in Wikidata and the world at large, so we feel that any cleaning process would introduce further biases. An instructive illustration of this is that the "gender" property in Wikidata is actually labeled in English as "sex or gender" (no distinction), and not limited to any value. Over our time snapshotting we found 36 values used for "sex or gender", including "male" and "female", but extending to non-binary genders "transgender female", "intersex", "fa'afafine", "transgender", "Gender fluid", "genderqueer", "kathoey", and "queer". At times the other categories of information are recorded here - perhaps erroneously - such as "gay", or "homosexuality". Cleaning this data would be a disservice, we feel, to communicating how Wikidata is used.

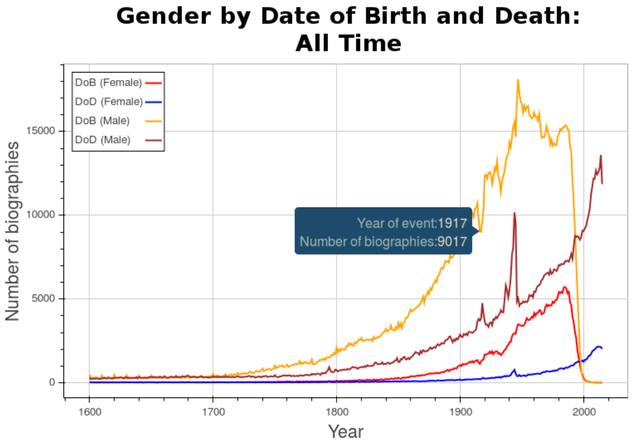


Figure 3: Wikidata gendered biographies aggregated by date of birth and date of death. We can see a noticeable spike in death for men around World War II, and that births drop about 20 years before the current year, as younger people tend not to be notable.

Our first snapshot is from September 17th 2014, and tracks the official Wikidata data dumps, updating weekly. We archived the January 3rd 2016 version as a quality-checked, canonical version [4]. All our code to make this data and the analyses presented here using both *python-pandas* and *R* can be found in our github repository [5].

Note that the missing data in the first half of 2015 is the period in which automatic collection of these statistics was under construction.

4. RESULTS

4.1 RQ1: Real World Validation of WHGI

This project is not the first to establish that Wikipedia data has real-world biases via correlation to exogenous indicators. In 2008 Rask concluded that Wikipedia editions displayed real-world *socio-economic* disparities similar to those found in the United Nations Human Development Index (HDI) [21]. However Rask identified three limitations with their study (1) they used only 11 Wikipedia languages (2) they suggested longitudinal analysis be carried out, and (3) language-disaggregation is only a proxy for country-disaggregation which the HDI utilises. WHGI overcomes all three of these limitations as it measures all 285 Wikipedia languages, updates weekly, and handles by-country-disaggregation. Although we are not focused on *socio-economic* disparities but *gender* disparities we still investigate exactly how well WHGI reflects the real world, and validated our indicators by comparing it against 3 exogenous datasets. We correlated the WHGI by date of birth versus historical world population trends; WHGI by country versus global gender-disparity indices; and WHGI by occupation versus United States Bureau of Labor Statistics occupation gender.

4.1.1 World Population

[3] based on the Inglehart-Welzel cultural map of the world

[4] https://figshare.com/articles/Wikidata_Human_Gender_Indicators/3100903

[5] https://github.com/notconfusing/WIGI

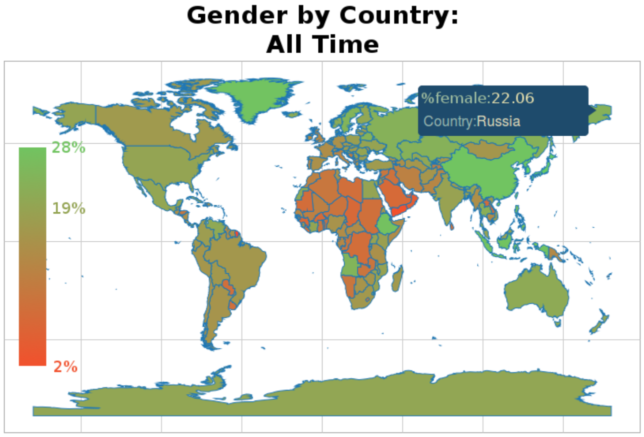


Figure 4: The Wikidata female ratio of biographies aggregated by place of birth and citizenship. The greener colours indicate a higher ratio of humans that are born-in or are citizens-of that country who are women. For instance China, Korea and Japan each approach 28% of biographies about women, whereas Russia and the U.S. are at 22% and 19% respectively.

Table 2: Correlation of humans in WHGI by date of birth and world population. $^{**}p \leq 0.01$**.**

snapshot	Pearson correlation
2014-09-17	0.852**
2016-01-03	0.845**

Our first validation is a "sanity check" to compare the world's population by year to the number of humans in WHGI by year of birth. We conduct this validation even though the number of people alive and the number of Wikipedia-notable people born are different measures. However if we operate under the assumptions that (a) the proportion of the world population which is Wikipedia-notable is constant over time and (b) that the birth rate is a fixed proportion of the population, then theoretically their curves should share approximately the same shape.

We performed a standard Pearson correlation between the number of people in Wikidata born in a particular year, and the estimated historical world population by the US Census Bureau[6]. The Census Bureau has estimates of world population from 10,000 BCE to 2001, and Wikidata from 4203 BCE to 2015. We conducted this correlation for our earliest and latest snapshots - the population statistics of Wikidata as of September 17th 2014, and again separately at January 3rd 2016. The results in Table 2 show a high and significant correlation between real world estimates and Wikidata, almost constant at 0.85. Overall though, the population of Wikidata over time seems very aligned with the World's population over time, so Wikidata at least is a "sane" representation of the world.

4.1.2 *Exogenous Gender-Disparity Indices*

WHGI is inspired, in part, by the rich landscape of gender disparity indices. This type of index ranks countries by some measure of gender disparity. If we aggregate WHGI by place of birth and citizenship, and look at the female ratio of humans, we also have a country-by-gender disparity measure[7]. We correlated the country rankings of this WHGI aggregation with four popular exogenous indices to see how well Wikidata reflects real world gender disparities.

Table 3: WHGI-country correlation to external indices. Correlation is the Spearman ρ**,** $^{*}p \leq 0.05$**,** $^{**}p \leq 0.01$**.**

snapshot	GEI	SIGI	GGGI	GDI
2014-09-17	0.417**	0.338**	0.310*	0.278**
2016-01-03	0.457**	0.402**	0.386**	0.299**

The four exogenous indices we used were: The traditional United Nations' Gender Development Index (**GDI**) [8] which considers disparity in income, education, and life expectancy. Social Watch's Gender Equity Index (**GEI**) [9] tries to broaden the scope of the variables by not only incorporating education and economic participation, but also by stretching into economic and political empowerment. The Global Gender Gap Index (**GGGI**) [10] grows yet wider by covering all previous topics but with more detail. And most recently the Social Institutions and Gender Index (**SIGI**)[11] has attempted to capture disparity in norms, values and attitudes.

For each index we conducted a calibration step, to find the date of birth start and end inclusion years which maximized our correlations. In each case the maximizing start year was found to be between 1900 and 1910 and the end year to be 2015. We interpreted the found thresholds as a good sign firstly because the exogenous indices are measures of recent history too, and secondly because it shows a robustness in the way that WHGI relates to exogenous indices.

We repeated the index correlations twice, once using the September 17th 2014 snapshot of Wikidata, and then using January 2016 data. That is, we have correlations between rankings of countries given by exogenous indices and Wikidata - at two separate times.

Table 3 shows the correlations with each index, all of which were significant and ranged from 0.278 to 0.457. Affirmingly, when looking at this information through a longitudinal lens, the correlation with every index is increasing over time at which we sampled Wikidata. On the low end the GDI correlation grew by 7.6%, and on the upper end, the SIGI correlation jumped 24.5% in the about-a-year time frame between Wikidata snapshots. Because we are using the female ratio of biographies by country and not the absolute number of biographies these correlations are not growing simply because of increased number of data points.

The WHGI being most closely correlated with GEI, and least to GDI has implications for Wikipedia's notability policy. Where they both measure gender gap in school enrollment, years of schooling, and earned income, GEI additionally measures positions of power, and GDI additionally measures life expectancy. That means that notability in Wikipedia is more correlated to power in society than it is to health status. As the strengths of these correlations has increased across all indices, and the order remains unchanged, this means that the gender disparities found in WHGI by country are increasingly looking *more* like these real world gender disparities.

4.1.3 *Occupation Gender*

The notion of what a human's job or occupation is, we see in Ta-

[6] https://commons.Wikipedia.org/wiki/File:Population_curve.svg

[7] Despite having the same by-country unit of analysis with this aggregation WHGI is not an "index" like those we compare it to, since an index weights and combines many indicators [23].

[8] http://hdr.undp.org/en/content/gender-development-index-gdi

[9] http://www.socialwatch.org/node/14366

[10] http://reports.weforum.org/global-gender-gap-report-2014/

[11] http://www.genderindex.org/ranking

Table 4: Rank correlation of gender ratios by occupation between WHGI and US Bureau of Labor Statistics. $^{**}p \leq 0.01$.

snapshot	Spearman Rank Correlation
2015-08-09	0.410**
2016-01-03	0.473**

ble 5, is well recorded in Wikidata at 58.7%. To answer the question of how representative of the real world Wikidata's gender by occupation is, we compared it to data from the United States Bureau of Labor Statistics (BLS)[12], borrowing this ground truth technique from [16] who used it to evaluate the gender representation of Google image search results for occupation key terms.

Approximately 60% of our sample have occupation data, and together over 4,000 occupations are represented. The BLS has 332 occupation categories which are at a higher level ontologically than recorded in Wikidata. Whereas Wikidata might record that someone is a pastry chef, the BLS only has a category for cooks. In order to match the indicators we used Wikidata's internal ontology hierarchy to generalize the occupation terms. A *subclass of* property exists in Wikidata, that relates items to their more general concept - which we can use for occupations. For instance Wikidata describes that pastry chef is a *subclass of* chef, and that chef is a *subclass* of cook.

Our method was to raise the generality of Wikidata occupations until there were less than 500 occupations to ease the matching task. Two authors then matched occupations manually for accuracy and confirmation. We resolved disagreements until the sets were matched. However not all occupations could be matched due to the specificity of the BLS, rendering coverage of Wikidata occupations 57% complete. The largest occupations in Wikidata were sportsperson and politician, and neither of them had matches in the BLS. In the reverse, there were many BLS occupations for which Wikidata did not have any matching occupations, such as "lodgings manager". This outlines a limitation of this validation, that being a lodgings manager does not inherently make you notable for inclusion in Wikipedia.

Finally we correlated the rankings of the list of most gendered occupations according to WHGI to that of the BLS. We did this for early and late snapshots, but because occupation was not a property that we initially recorded, our first snapshot which included occupation was August 9th 2015. Table 4 shows the Spearman rank correlation found then was 0.410, and since the correlation has increased to 0.473. These are moderate correlations which we claim support a link that Wikidata reflects the real world, and like the by-country indicator is becoming increasingly more accurate.

It must be acknowledged also that the BLS data describes the United States whereas WHGI has a worldwide scope. However when we restricted WHGI to only biographies with place of birth in the United States the Spearman rank correlation became only marginally significant. This might be due to the fact that the population in Wikidata that has both occupation and place of birth recorded are particularly notable people, and contain even less of the everyday occupations that the BLS does.

4.2 RQ2: Relation to Existing Editing Efforts

Another main purpose for investigating this dataset is to support and provide metrics for Wikipedian communities attempting to address content gaps. Therefore we turn to focus on statistics of our dataset with regard to how it has changed over time as these Wikipedian communities have been editing.

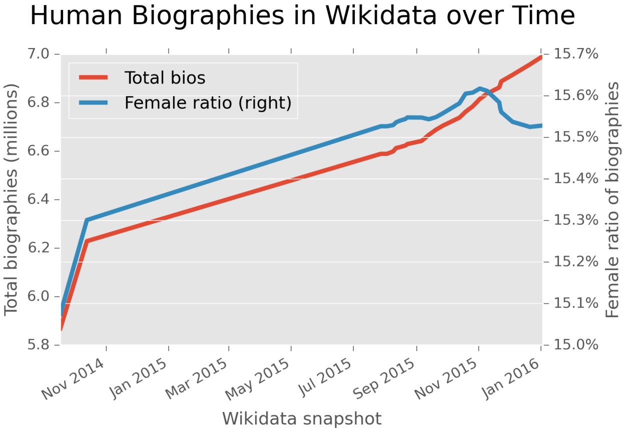


Figure 5: Total number of human biographies (left) and the female ratio of those biographies (right) by Wikidata snapshot. The total number of humans found in Wikidata over time is displaying linear, unconstrained growth. Over the same time the female ratio of biographies in Wikidata has risen by 0.5%.

First we queried the way the total number of biographies and the ratio of women represented as Wikidata has evolved. During our observation total humans in Wikidata increased from 5,869,606 to 6,999,542, and shows linear, unconstrained growth (see Figure 5). These may be thought of as the total biographies across all Wikipedias – although about 1-2% of these humans exist only in Wikidata and not in any Wikipedia. An important measure for content-focused communities is the ratio of biographies which are about women, as espoused in *WikiProject Women in Red*[13]. We looked into the ratio of humans recorded "female" versus all gendered biographies. Similar to total biographies this measure is rising at a fairly linear rate of approximately 0.5% per year (Figure 5). The final months on record show a slight decline which warrants further investigation.

We took a more granular look at the evolution of the female ratio of biographies by disaggregating by Wikipedia Language in Figure 6. Of the languages that have 100,000 or more gendered biographies, during our observation period all Wikis increased in total biographies. In terms of the rate of women represented in biographies, Norwegian (Bokmål), Spanish, English, Finnish, and Dutch Wikipedias each increased by more than 0.5% points. Japanese however jumped the most, by 4% points, more than any other language.

Japanese Wikipedians relate this increase to strong editing activity about women who are idols, models and celebrities, despite an effort to delete biographies about Adult Video idols[14]. We hypothesize this deletion effort may also be partially responsible for the decline viewable in Chinese Wikipedia and the decline in total ratio seen in Figure 5.

Are these changes in Wikipedias correlated with the efforts of Wikipedian communities targeting gendered content? There are many Wikipedian communities who's goal it is is to increase the coverage of women's biographies, for instance: WikiProject Women

[12] http://www.bls.gov/cps/aa2012/cpsaat11.htm

[13] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red

[14] https://twitter.com/halowand/status/712636154642706433

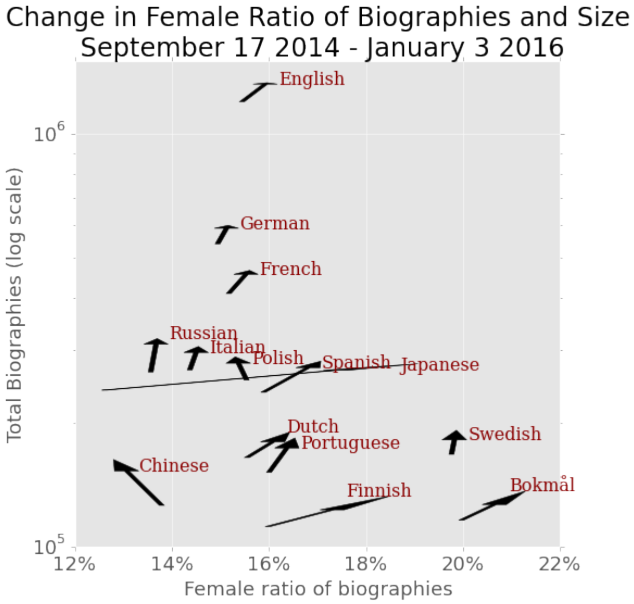


Figure 6: The change in female ratio of biographies and total biographies over time, of Wikipedia languages with 100,000 or more gendered biographies. The tail of the arrow represents the Wikipedia's position in September 2014, and the head in January 2016.

Scientists [15], Art + Feminism[16], and Women in Red (see a more complete list at [17]). One concern of these organizations is if their editing efforts are making large scale impacts on the Wikipedia. Fortunately Wikiproject Women in Red keeps metrics of how many biographies they add on a monthly basis, and we were able to conduct a cross-correlation between the time-series of monthly number of biographies created by Women in Red, and the number of female biographies added to English Wikipedia. The correlation between these activities is 0.657, which indicates a positive link. We cannot determine a causal relationship between these two variables – it may just be due to a general trend. Even so we are able to numerically highlight a link between editing efforts and the increase in women's representation in English Wikipedia which was not viewable before.

4.2.1 Data Quality

Of course the female ratio of biographies is not the only way to characterize the effect of content-focused editing, we might also inquire to the wider quality of biographies in Wikidata. One way to investigate data quality is the coverage of demographic properties of these biographies. Figure 7 shows the trend in coverage of all properties at the earliest and latest WHGI snapshots and Table 5 compares this coverage to latest DBpedia snapshot[18] - an independent project to extract data from Wikipedia. The statistics show that data quality has been increasing across all properties over time. The number of humans with *gender* data increased just 1% point but is close to complete at 96% coverage. In the time domain however *date of birth* and *date of death* coverage increased by 14% points and 7% points respectively. This means that date of birth and death cover about $\frac{3}{4}$ and $\frac{2}{3}$, of the biographies, both counts of which exceed DBpedia which count less than $\frac{1}{2}$ and $\frac{1}{4}$ respectively. *Place of birth* jumped by 6% points to $\frac{1}{3}$ coverage, however this lags behind DBpedia's strength in this domain at $\frac{4}{5}$.

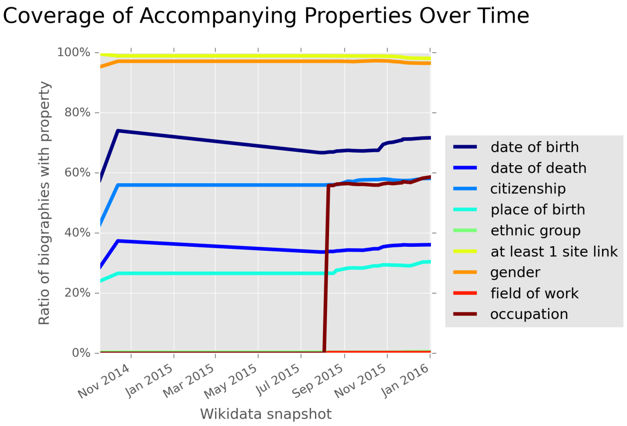


Figure 7: Trends of property coverage by Wikidata snapshot. Most humans have at least one Wikipedia article ("site link") and a recorded gender, other properties are slowly increasing in coverage.

Table 5: Change in rates of property coverage for humans between earliest and latest snapshots and DBpedia 2015-10

	2014-09-17	2016-01-03	DBpedia
2015-10			
gender	95.3%	96.5%	n/a
date of birth	57.6%	71.7%	45.3%
date of death	28.6%	36.1%	17.8%
citizenship	42.8%	58.2%	n/a
place of birth	24.0%	30.5%	80.0%
ethnic group	0.3%	0.6%	n/a
field of work	n/a	0.3%	n/a
occupation	n/a	58.7%	n/a
at least 1 site link	99.6%	98.1%	n/a

In terms of Wikidata specific property coverage *Citizenship* data increased the most, by 15% points and is available for more than half of humans, Next *ethnic group* doubled, but just to 0.6% coverage. Unfortunately *field of work*, and *occupation* data was not included in our dataset until later, so their growth, while increasing, is not precisely comparable.

Curiously, the rate of humans having at least one Wikipedia article decreased slightly, and this has an important interpretation. A Wikidata human without a Wikipedia article is known as a "structural item", for instance a member of royalty without a Wikipedia article but is needed to make a family tree complete. With the view that a structural item is an artefact from editors paying attention to Wikidata's structure, the decrease in the ratio of sitelinked humans can also be seen as an increase in data quality.

5. DISCUSSION

We established the WHGI, and using longitudinal analysis showed

[15] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_scientists

[16] http://art.plusfeminism.org/about/

[17] https://meta.wikimedia.org/wiki/Address_the_gender_gap/Initiatives

[18] http://wiki.dbpedia.org/services-resources/datasets/dataset-2015-10/dataset-2015-10-statistics

that women are increasingly being represented in most Wikipedias. This, along with the fact that exogenous validations show that Wikipedias' gender disparities are increasingly reflecting the real world suggest that Wikipedia is "catching up" to real world disparities. Further, this catching-up is also correlated to the activity levels of women-focused editing initiatives.

But as well as providing metrics for Wikipedian communities to monitor and potentially address the content gap, WHGI can also be used in collaboration with other quality measures to impact research, particularly to enhance our understanding of the gender disparity phenomenon in Wikipedia and shed light on the underlying mechanisms in peer production communities.

For instance Warncke-Wang et al. explored supply and demand misalignment in Wikipedia by looking at the difference in actual and predicted article quality using page-views [27]. They cite examples of over-demanded articles where readership exceeds quality (e.g. "Wedding", and "cisgender") and over-supplied articles where quality exceeds readership (e.g. "Themes in Robert Browning's poetry"). Following these examples, they follow up with a broader investigation of patterns in misalignment by using WikiProjects to categorize articles. For example, they find that the reader demand of WikiProject:LGBT (Lesbian, Gay, Bisexual, Trans*) far exceeds Wikipedia's supply: articles in Wikiproject LGBT are 9 times more likely to be found in the "Needs Improvement" dataset compared to the project's overall representation in Wikipedia. They call this measure the *relative risk*[4].

We can further our understanding of the extent and nature of gender disparity in Wikipedia by combining our WHGI metrics and Warncke-Wang's supply-demand misalignment approach. We started with the same dataset in Warncke-Wang et al's paper and reran a variant of their analysis. Instead of aggregating by WikiProject, we utilised the WHGI and were able to aggregate by gender over the entire encyclopedia. Table 6 shows the relative risk for gendered biographies being either in need of improvement (over-demanded), or showing extra "spent effort" (over-supplied). Our findings show that women's biographies are 30% more likely to be in the category of "needs improvement", whereas men's biographies are 14% less likely. Women's biographies are roughly comparable to the average encyclopedia article in terms of belonging to the "spent effort" category, whereas men's biographies show a 15% higher likelihood.

The row "non-binary" speaks to the 152 biographies whose gender is neither "male" nor "female", and shows twice the likelihood to need improvement, $\frac{1}{2}$ the likelihood to represent spent effort. As a control, we also calculated the relative risk of every article which is not a biography, in the "non-biography" row, which sensibly shows a close-to-baseline result in all categories.

Another example of how WHGI could be useful for researchers is that the time series nature of the WHGI dataset enables an alternative outcome metric to measure the efficiency of a range of content improvement movements and strategies in Wikipedia. Researchers have long been interested in understanding the effectiveness of content-improvement movements, ranging from formally organized efforts such as *APS Wikipedia* initiatives [7] and the *Wikipedia Education Program* [26], to members self-organizing activities such as *Collaboration of the Week*, *WikiCup*, and *Today's Article for Improvement* [29][26]. To measure the success of these movements, prior research has used outcome metrics such as number of revisions [29][7], and Wikipedia's article quality measurement and machine predicted quality measurement [26]. WHGI can provide a new success metric - the gender ratio of the biographic articles created by a particular movement - to provide a new angle to examine the efficiency of these movements in improving the content coverage in Wikipedia.

Table 6: The relative risk for biography article quality to be misaligned with readership demand in English Wikipedia. Biographies about women are 30% more likely to have readership exceed article quality, whereas biographies about men are 14% less likely to need improvement. Spent Effort cases are where article quality exceeds readership demand, and men's more than women's biographies display this mismatch.

	Needs Improvement	Spent Effort
Female	1.30	1.03
Male	0.86	1.15
Non-binary gender	2.65	0.50
Non-biography	1.03	0.95

Monitoring the effect of initiatives provokes asking about the gendered implications of other environmental factors such as *notability policies* which could be tested using WHGI and propensity score matching. That is, for a language which has implemented a stricter notability policy we could compare the way that language's gender composition changes against a language which was on a similar trajectory before the change, but retained its notability policy.

Beyond the Wikipedia-research landscape, a historian could use the data to determine the gender-disparity levels of a specific place and time. Typically to quantify the gender climate one would rely on the indices like those mentioned in the exogenous indices section. However these indices, are limited to discussing recent history. Our validation showed that our data is in touch with the real world. With this dataset we can quantify a type of gender-disparity of medieval France, ancient Greece, or Ming Dynasty China. WHGI is useful in all the same ways that exogenous indices are used, only with a larger time-span. That is an approach not possible before Wikidata.

6. LIMITATIONS AND FUTURE WORK

WHGI is a proxy for real world phenomena, but is also limited by the worldview of Wikipedia editors and is constrained by its notability policies. Wikipedia's notability policies generally require humans to be in positions of power – although some exceptions may exist, such as notable victims etc. Requiring positions of power for notability systemically biases inclusion against women. How then can the general rise in women's biographical representation be explained? There could be several possible reasons. At least three factors that affect encyclopedic inclusion are: (1) the rate at which women receive positions of power in the real world, (2) the level of gender bias in Wikipedias' notability policies, and (3) the level of efforts to write about women in Wikipedia.

The validation analysis (correlations with existing gender disparity indices) show that WHGI captures real world phenomena. Since we were able to look at the intra-year level, and assuming that real world disparities are less less fast-moving, the increased correlation with existing gender disparity indices between our earliest and latest snapshots indicates that WHGI is capturing movement in editing effort and policy changes in the Wikipedia community. Unfortunately we still cannot disentangle the influence of these factors.

Yet another limitation is that there may also be biography articles in Wikipedias that are not recorded as humans in Wikidata. However the size of this set is not readily computable, and our best estimates come from the latest DBpedia extract which contains

3,158,512 humans - 122,276 or 4% more than Wikidata. Particularly we would like to know how much growth in Wikidata stems from Wikipedia's growth (e.g. a new biography is added), or migration of information to Wikidata (e.g. an existing biography in Wikipedia is marked as a human in Wikidata).

Still applications on top of WHGI indicators could be built to help the community, for instance to detect spikes in creation and deletion of specific demographics of humans. Such a tool could also alert to the presence of unplanned activity, good or bad, which affects the macro-level gender of Wikipedia and Wikidata. Take Figure 1, it shows a week where contributions to Nepali Wikipedia are nearly 100% about women. Likewise, if a week were to show a net-subtraction of female biographies a community alert could be generated.

We wonder if tools like this would be motivating as well, as well as simply measuring progress. This would suggest a path of research investigating how editors use information to act. As it happens Women in Red already cite WHGI for data, but would a tighter integration help the cause? A user study in providing editing communities with detailed feedback and exact movement in the landscape would ask if "what gets measured, get fixed" - more quickly?

7. CONCLUSION

We made the Wikidata Human Gender Indicators (WHGI), a biographic database for researchers wishing to incorporate gender data along dimensions of time, space and occupation. Based on Wikidata and Wikipedia it can most obviously be used by those communities to monitor the effects of focused editing and biases in their content. We also validate the indicators with measures of the real world, such as population, country-based gender disparities, and occupations. These validations show that the WHGI is significantly correlated to real world demographics and gender disparities. In addition, we demonstrate that data quality of Wikidata has been increasing. Data quality and correlations increasing together is particularly encouraging as support for using WHGI as a tool, as we example by extending previous research in article-readership misalignment. WHGI is freely available for download, we outline some of the potential ways in which it could be used, and hope that many more are thought of by others.

8. ACKNOWLEDGMENTS

We acknowledge the hard work done by Wikimedians in countering systemic bias. We hope to continue running the open source project in service of the Wikipedia and research communities seeking to statistically describe gender disparities. The ways in which we expand WHGI we hope will be directed by user's feedback.

We are especially grateful to the Wikimedia Foundation for funding us through an Individual Engagement Grant.

We are also especially grateful to the Wikidata and Wikidata Toolkit teams.

9. REFERENCES

[1] S. Burgstaller-Muehlbacher, A. Waagmeester, E. Mitraka, J. Turner, T. E. Putman, J. Leong, P. Pavlidis, L. Schriml, B. M. Good, and A. I. Su. Wikidata as a semantic framework for the Gene Wiki initiative. *bioRxiv*, page 032144, Nov. 2015.

[2] P. Burstein. *Equal Employment Opportunity: Labor Market Discrimination and Public Policy*. Transaction Publishers, 1994.

[3] N. Cohen. Wikipedia Ponders Its Gender-Skewed Contributions. *The New York Times*, Jan. 2011.

[4] H. T. O. Davies, I. K. Crombie, and M. Tavakoli. When can odds ratios mislead? *BMJ*, 316(7136):989–991, Mar. 1998.

[5] S. Eckert and L. Steiner. (Re)triggering Backlash: Responses to News About Wikipedia's Gender Gap. *Journal of Communication Inquiry*, 37(4):284–303, Oct. 2013.

[6] Y.-H. Eom, P. Aragón, D. Laniado, A. Kaltenbrunner, S. Vigna, and D. L. Shepelyansky. Interactions of cultures and top people of wikipedia from ranking of 24 language editions. *PLoS ONE*, 10(3):e0114825, 03 2015.

[7] R. Farzan and R. E. Kraut. Wikipedia Classroom Experiment: Bidirectional Benefits of Students' Engagement in Online Production Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 783–792, New York, NY, USA, 2013. ACM.

[8] R. Glott, P. Schmidt, and R. Ghosh. Wikipedia survey – overview of results. *United Nations University: Collaborative Creativity Group*, 2010.

[9] D. Goldfarb, D. Merkl, and M. Schich. Quantifying Cultural Histories via Person Networks in Wikipedia. *arXiv:1506.06580 [physics]*, June 2015. arXiv: 1506.06580.

[10] E. Graells-Garrido, M. Lalmas, and F. Menczer. First Women, Second Sex: Gender Bias in Wikipedia. *arXiv:1502.02341 [cs]*, pages 165–174, 2015. arXiv: 1502.02341.

[11] S. A. Hale. Multilinguals and Wikipedia Editing. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 99–108, New York, NY, USA, 2014. ACM.

[12] E. Hargittai and A. Shaw. Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia. *Information, Communication & Society*, 18(4):424–442, Apr. 2015.

[13] C. Heward and S. Bunwaree. *Gender, Education and Development: Beyond Access to Empowerment*. Palgrave Macmillan, Feb. 1999.

[14] B. M. Hill and A. Shaw. The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLoS ONE*, 8(6):e65782, June 2013.

[15] M. Hinnosaar. Gender Inequality in New Media: Evidence from Wikipedia. SSRN Scholarly Paper ID 2617021, Social Science Research Network, Rochester, NY, May 2015.

[16] M. Kay, C. Matuszek, and S. A. Munson. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. pages 3819–3828. ACM Press, 2015.

[17] S. Klasen. Gender-Related Indicators of Well-Being. Technical Report 102, Discussion Papers / Universität Göttingen, Ibero-Amerika-Institut für Wirtschaftsforschung, 2004.

[18] M. Krötzsch. How to use Wikidata: Things to make and do with 40 million statements, 2014.

[19] S. T. K. Lam, A. Uduwage, Z. Dong, S. Sen, D. R. Musicant, L. Terveen, and J. Riedl. WP:Clubhouse?: An Exploration of Wikipedia's Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 1–10, New York, NY, USA, 2011. ACM.

[20] E. Mitraka, A. Waagmeester, S. Burgstaller-Muehlbacher, L. M. Schriml, A. I. Su, and B. M. Good. Wikidata: A

platform for data integration and dissemination for the life sciences and beyond. *bioRxiv*, page 031971, Nov. 2015.

[21] M. Rask. The reach and richness of Wikipedia: Is Wikinomics only for rich countries? *First Monday*, 13(6), May 2008.

[22] J. Reagle and L. Rhue. Gender Bias in Wikipedia and Britannica. *International Journal of Communication*, 5(0):21, Aug. 2011.

[23] R. J. Rossi and K. J. Gilmartin. *The handbook of social indicators: sources, characteristics, and analysis*. Garland STPM Press, 1980.

[24] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):78–85, Sept. 2014.

[25] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Ninth International AAAI Conference on Web and Social Media*, Apr. 2015.

[26] M. Warncke-Wang, V. R. Ayukaev, B. Hecht, and L. G. Terveen. The Success and Failure of Quality Improvement Projects in Peer Production Communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 743–756, New York, NY, USA, 2015. ACM.

[27] M. Warncke-Wang, V. Ranjan, L. Terveen, and B. Hecht. Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. In *ICWSM 2015: Ninth International AAAI Conference on Web and Social Media*, 2015.

[28] World Health Organization, editor. *Women and health: today's evidence tomorrow's agenda*. World Organization, Geneva, 2009.

[29] H. Zhu, R. Kraut, and A. Kittur. Organizing Without Formal Organization: Group Identification, Goal Setting and Social Modeling in Directing Online Production. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 935–944, New York, NY, USA, 2012. ACM.